Chromatin-accessibility estimation from single-cell ATAC data with scOpen

Zhijian Li[1, +], Christoph Kuppe[2,3 +], Susanne Ziegler[2], Mingbo Cheng[1], Nazanin Kabgani[3], Sylvia Menzel[2], Martin Zenke[4, 5], Rafael Kramann[2,3,6 *], and Ivan G. Costa[1, *]

[1]Institute for Computational Genomics, Joint Research Center for Computational Biomedicine, RWTH Aachen University Medical School, 52074 Aachen, Germany

[2]Institute of Experimental Medicine and Systems Biology, RWTH Aachen University Medical School, 52074 Aachen, Germany

[3]Division of Nephrology and Clinical Immunology, RWTH Aachen University, 52074 Aachen, Germany

[4]Department of Cell Biology, Institute of Biomedical Engineering, RWTH Aachen University Medical School, Aachen, 52074, Germany

[5]Helmholtz Institute for Biomedical Engineering, RWTH Aachen University, Aachen, Germany

[6]Department of Internal Medicine, Nephrology and Transplantation, Erasmus Medical Center, 3015GD Rotterdam, The Netherlands

[*]corresponding authors: rkramann@ukaachen.de, ivan.costa@rwth-aachen.de

[+]these authors contributed equally to this work

Abstract

A major drawback of single cell ATAC (scATAC) is its sparsity, i.e. open chromatin regions with no reads due to loss of DNA material during the scATAC-seq protocol. We propose scOpen, a computational method for imputing and quantifying the open chromatin status of regulatory regions from sparse scATAC-seq experiments. We show that scOpen improves crucial down-stream analysis steps of scATAC-seq data as clustering, visualisation, *cis*-regulatory DNA interactions and delineation of regulatory features. We demonstrate the power of scOpen to dissect regulatory changes in the development of fibrosis in the kidney. This identified a novel role of Runx1 and target genes by promoting fibroblast to myofibroblast differentiation driving kidney fibrosis.

Introduction

The simplicity and low cell number requirements of assay for transposase-accessible chromatin using sequencing (ATAC-seq)[1] made it the standard method for detection of open chromatin enabling the first study of open chromatin of cancer cohorts[2]. Moreover, careful consideration of digestion events by the enzyme Tn5 allowed insights on regulatory elements as positions of nucleosomes[1,3], transcription factor binding sites and the activity level of transcription factors[4]. The combination of ATAC-seq with single cell sequencing (scATAC-seq)[5] further expanded ATAC-seq applications by measuring the open chromatin status of thousands of single cells from healthy[6,7] and diseased tissues[8]. Computational tasks for analysis of scATAC-seq include detection of novel cell types with clustering (scABC[9], cisTopic[10], SnapATAC[11]); identification of transcription factors (TF) regulating individual cells (chromVAR[12]); and prediction of co-accessible DNA regions in groups of cells (Cicero[13]).

Usually, the first step for analysis of scATAC-seq is detection of open chromatin regions by calling peaks on the scATAC-seq library by ignoring cell information. Next, a matrix is built by counting the number of digestion events per cell in each of the previously detected regions. This matrix usually has a very high dimension (up to $> 10^6$ regions) and a maximum of two digestion events are expected for a region per cell. As with scRNA-seq[14–16], scATAC-seq is effected by dropout events due to loss of DNA material during library preparation. These characteristics render scATAC-seq count matrix extremely sparse, i.e. 3% of non-zero entries. In contrast, scRNA-seq have less severe sparsity (> 10% of non-zeros) than scATAC-seq due to smaller dimension (<20.000 genes for mammalian genomes) and lower dropout rates for genes with high or moderate expression levels. This sparsity poses challenges in the identification of cell specific open chromatin regions and is likely to affect downstream analysis as clustering and detection of regulatory features. So far, only few computational approaches addresses the extreme sparsity and propose imputation methods for scATAC-seq data[10,17].

We here present scOpen, which is an unsupervised model for scATAC-seq imputation. It estimates accessibility scores to indicate if a region is open in a particular cell. The imputed matrix can be used as input for usual computational methods of scATAC-seq data as clustering, visualisation and prediciton of DNA-interactions (**Fig. 1a**). We demonstrate the power of scOpen on a comprehensive benchmarking analysis using publicly available scATAC-seq data with true labels. Moreover, we use scOpen together with HINT-ATAC[4] footpriting analysis to infer regulatory networks driving the development of fibrosis with a novel scATAC-seq time-course dataset of 31,000 cells in murine kidney fibrosis, identifying Runx1 as a novel regulator of myofibroblast differentiation.

Results

Open chromatin estimation with scOpen

scOpen performs imputation and denoising of a scATAC-seq matrix via a regularised non-negative matrix factorisation (NMF) based on a binarised scATAC-seq cell count matrix, where features represent open chromatin (OC)

regions which are obtained by peak calling based on aggregated scATAC-seq profiles. This matrix is transformed using term frequency–inverse document frequency (TF-IDF), which weights the importance of an OC region to a cell. Next, it applies a regularised NMF using a coordinate descent algorithm[18]. In addition, it provides a computational approach to optimise the dataset specific rank k of the NMF approach based on a knee detection method[19]. scOpen provides as results imputed and reduced dimension matrices, which can be used for distinct downstream analysis as visualisation, clustering, inference of regulatory players and *cis*-regulatory DNA interactions (**Fig .1a**).

First, we made use of simulated scATAC-seq similar as in Chen et al., 2019[20] to evaluate the parametrisation of two hyper-parameters of scOpen, i.e., the rank k and the regularisation term λ (see Methods and **Supplementary Fig. 1a-d**). Results indicate that scOpen automatic procedure for rank selection obtains close to optimal results, i.e. selected rank had similar accuracy than best ranks for both imputation and clustering problems. Regarding λ, a value of 1 is optimal in the imputation problem, where values in the range $[0, 1]$ were optimal for the clustering problem. This indicates the importance of the regularisation parameter in scATAC-seq data imputation. The λ=1 and the rank selection strategy are used as default by scOpen.

Benchmarking of scOpen for imputation of scATAC-seq

For benchmarking, we made use of four public scATAC-seq data sets: cell lines[5], human hematopoiesis composing of eight cell types[6], four sub-types of T cells[8], and a multi-omics RNA-ATAC from peripheral blood mononuclear cells (PBMCs) with fourteen cell types (Methods). These datasets were selected due to the presence of external labels, which are used to estimate evaluation measures. After processing, we generated a count matrix for each dataset and detected 50k to 120k open chromatin regions with 3-7% of non-zero entries, confirming the extreme sparsity of scATAC-seq data (**Supplementary Table 1**). For comparison, we selected top performing imputation/denoising methods[21] proposed for scRNA-seq (MAGIC[14], SAVER[22], scImpute[23], DCA[24] and scBFA[25]); two scATAC-seq imputation methods (cisTopic[10] and SCALE[17]); one control method (PCA[26]); and the raw count matrix (**Supplementary Fig. 2a**).

We first evaluated time and memory requirement of imputation methods (Methods). scOpen had overall lowest memory requirements, i.e it required at least 2 fold less memory as compared to cisTopic, MAGIC or SCALE (**Fig. 1b**) and had a maximum requirement of 16GB on the PBMC dataset (**Supplementary File 1**). Regarding computing time, MAGIC was the fastest followed by SCALE and scOpen. These were the only methods performing the imputation of the large PBMCs dataset (10k cells vs. 100k peaks) in less than 3 hours (**Fig. 1c**), while PCA, Saver and DCA failed to execute at the PBMCs dataset.

We next tested if imputation methods can improve recovery of true OC regions. For this, we created true/negative OC labels for each cell type by peak calling of bulk ATAC-seq profiles. Next, we evaluated the correspondence between imputed scATAC-seq values and peaks of the corresponding cell type with the area under precision recall curve (AUPR) (Methods). scOpen significantly outperformed all competing methods by presenting the highest mean

AUPR (**Fig. 1d**). The combined ranking indicates SCALE and MAGIC as runner up methods (**Supplementary Fig. 2b**).

We also investigated the impact of imputation on the estimation of distances between cells and the impact on standard clustering methods. Distance between cells were evaluated with the silhouette score, while clustering accuracy was evaluated with Adjusted Rand Index (ARI)[27] both regarding the agreement with known cell labels. scOpen was the best performer in all data sets regarding the silhouette score (**Fig. 1e**). The combined ranking demonstrated that scOpen had significantly better results than competing methods, while cisTopic and MAGIC were runner-up methods (**Supplementary Fig. 2c**). Regarding clustering, scOpen was best in the hematopoiesis and multi-omics PBMCs datasets and second best for cell lines and T cell datasets (**Fig. 1f**). When considering the combined ranking scOpen performed best (**Supplementary Fig. 2d**) followed by cisTopic and MAGIC. The discriminative power of scOpen was also supported by UMAP[28] projections of these datasets, which provide clear separation of the majority of cell labels (**Supplementary Fig. 3**). Altogether, these results support that scOpen outperforms state-of-the-art imputation methods, while providing the lowest memory footprint and an above average time performance.

Benchmarking of scATAC-seq clustering methods

Another relevant question was to compare scOpen with top performing state-of-the-art scATAC-seq pipelines: cisTopic, SnapATAC and Cusanovich2018[20] (Methods; **Supplementary Fig. 4a**). Here, pipelines were evaluated with the default clustering methods, i.e graph based clustering for SnapATAC[11] and density based clustering for other methods[10]. We also evaluated the use of both reduced and imputed matrices for scOpen and cisTopic, as these methods provide both type of representations.

The evaluation of distance matrices with the silhouette score indicated that both imputed or low dimension scOpen matrices presented the highest score in all data-sets (**Fig. 2a**) and both scOpen matrix representations tied as first in the combined rank (**Supplementary Fig. 4b**). cisTopic, which was the runner up method, performed well in cell lines, hematopoiesis and T-cells but poorly for multi-omics PBMCs. Next, we evaluated clustering performance of competing pipelines. Again, scOpen performed best on cell lines and hematopoiesis data sets and ranked first/second in the combined rank (**Supplementary Fig. 4c**). Overall, this analysis indicates that both reduced dimension and imputed scOpen matrices obtain best overall results for distance and clustering representations on evaluated datasets. Of note, the low dimensional matrix reduces memory footprint on the clustering by > 1000 fold in comparison to use of full imputed matrices serving as an alternative for clustering of large dimensional data sets.

Improving scATAC-seq downstream analysis using scOpen estimated matrix

Next, we tested the benefit of using scOpen estimated matrices as input for scATAC-seq computational pipelines, which have as objective the identification of regulatory features associated to single cells (chromVAR[12]), estimation

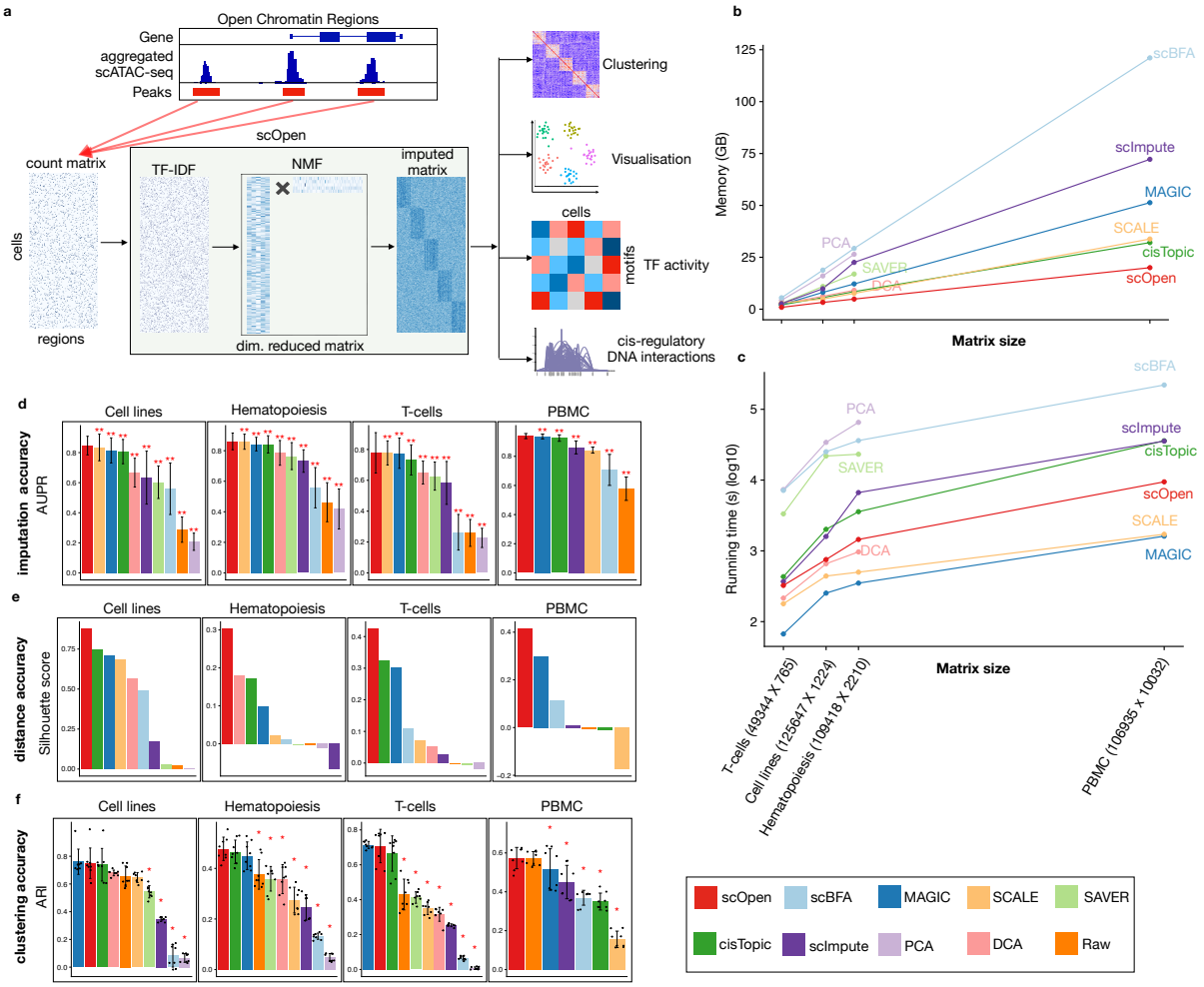

Fig. 1. scOpen and benchmarking of scATAC-seq imputation methods. a, scOpen receives as input a sparse peak by cell count matrix. After matrix binarisation, scOpen performs TF-IDF transformation followed by NMF for dimension reduction and matrix imputation. The imputed or reduced matrix can then be given as input for scATAC-seq methods for clustering, visualisation and interpretation of regulatory features. **b,** Memory requirements (y-axis) of imputation/denosing methods on benchmarking datasets (x-axis). The x-axis represents the number of elements of the input matrix (number of OC regions by cells). **c,** Same as **b** for running time requirements. **d,** Barplots showing evaluation of imputation/denosing methods for recovering true peaks. y-axis indicates mean AUPR of the cells by applying the methods to the scATAC-seq matrix from benchmarking datasets. Error bars represent standard deviation of AUPR. The asterisk and the two asterisks mean that the method is outperformed by the top ranked method (scOpen) with significance levels of 0.05 and 0.01, respectively. **e,** Barplots showing silhouette score (y-axis) for benchmarking datasets. **f,** Barplots showing the clustering accuracy for distinct imputation methods. y-axis indicates Adjusted Rand Index. Dots represent individual ARI values of distinct clustering methods. The asterisk and the two asterisks mean that the method is outperformed by the top ranked method with significance levels of 0.05 and 0.01, respectively.

of gene activity scores and DNA-interactions (Cicero[13]) or a clustering method tailored for scATAC-seq data (scABC[9]) (**Supplementary Fig. 4d**). Both chromVAR and Cicero first transform the scATAC-seq matrix to either transcription factors and genes feature spaces respectively. Clustering was then performed using the standard pipelines from each approach. We compared the clustering accuracy (ARI) and distance (silhouette score) of these methods with either raw or scOpen estimated matrices. In all combinations of methods and datasets, we observed

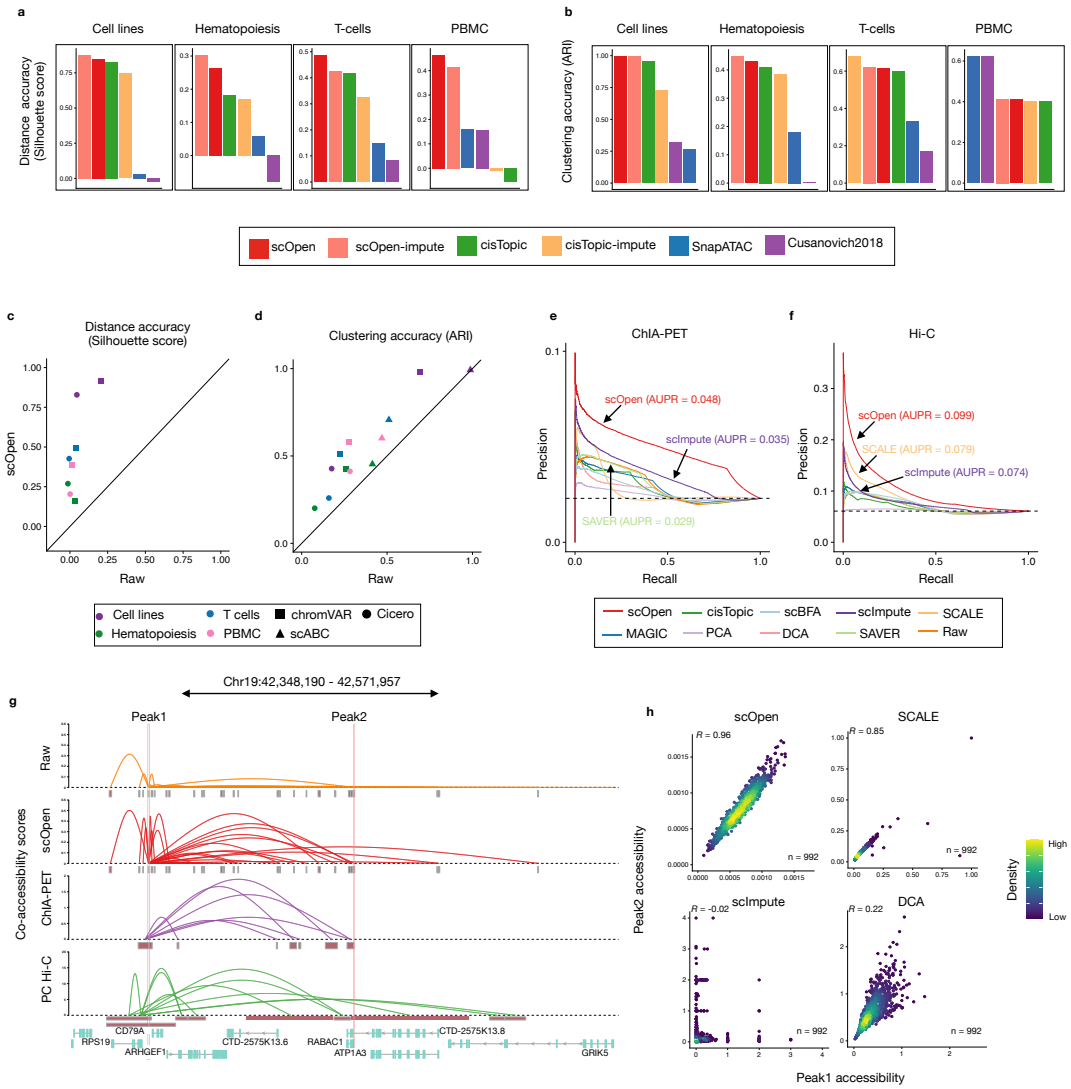

Fig. 2. Benchmarking of scATAC-seq clustering and downstream analysis a, Bar plots showing an evaluation of distances estimated on distinct scATAC-seq representations with a silhouette score. **b,** Barplots showing the clustering accuracy (ARI) for distinct clustering pipelines. **c,** Scatter plot comparing silhouette score of datasets by providing raw (x-axis) and scOpen estimated matrices (y-axis) as input for Cicero and chromVAR. Colours represent dataset and shapes represent methods. scABC is not evaluated as it does not provide a space transformation. **d,** Same as **c** for clustering results (ARI) of Cicero, chromVAR and scABC. **e,** Precision-recall curves showing evaluation of the predicted links on H1-ESC cells using raw and imputed matrix as input. We used data from pol-II ChIA-PET as true labels. Colours refer to methods. We reported the AUPR for top 3 methods. **f,** Same as **e** by using Hi-C data as true labels. **g,** Visualisation of co-accessibility scores (y-axis) of Cicero predicted with raw and scOpen estimated matrices contrasted with scores based on RNA pol-II ChIA-PET (purple) and promoter capture Hi-C (green) around the *CD79A* locus (x-axis). For ChIA-PET, the log-transformed frequencies of each interaction PET cluster represent co-accessibility scores, while the negative log-transformed p-values from the CHiCAGO software indicates Hi-C scores.**h,** Scatter plot showing single cell accessibility scores estimated by top-performing imputation methods (according to **f**) for the link between peak 1 and peak 2 (supported by Hi-C data). Each dot represent a cell and colour refers to density. Pearson correlation between is shown on the left-upper corner.

a higher or equal ARI/silhouette whenever a scOpen matrix was provided as input (**Fig. 2c-d**). These results were also reflected in the UMAP visualisation with and without scOpen imputation (**Supplementary Fig. 5**).

Prior to estimating gene centric open chromatin scores, Cicero first predicts co-accessible pairs of DNA regions in groups of cells, which potentially form *cis*-regulatory interactions. We compared Cicero predicted interactions on human lymphoblastoid cells (GM12878) by using Hi-C and ChIA-PET from this cell type as true labels for all imputation methods with data as provided in Pliner *et al. 2018*[13]. Both AUPR values and odds ratios indicated that the scOpen matrix improves the detection of GM12878 interactions globally (**Fig. 2e-f**; **Supplementary Fig. 6a-b**). The power of scOpen imputation was clear when checking the individual locus (**Fig. 2g**), as previously described by Cicero[13]. This is evident when contrasting accessibility scores between pairs of peak-to-peak links supported by Hi-C predictions (**Fig. 2h; Supplementary Fig. 6c-g**). scOpen obtained highly correlated accessibility scores, while other imputation methods showed quite diverse association patterns. Together, these results indicated that the use of scOpen estimated matrices improves downstream analysis of state-of-the-art scATAC-seq methods.

Applying scOpen to scATAC-seq of fibrosis driving cells

Next, we evaluated scOpen in its power to improve detection of cells in a complex disease dataset. For this, we performed whole mouse kidney scATAC-seq in C57Bl6/WT mice in homeostasis (day 0) and at two time points after injury with fibrosis: 2 days and 10 days after unilateral ureteral obstruction (UUO)[29,30]. Experiments recovered a total of 30,129 high quality cells after quality control with average 13,933 fragments per cell, a fraction of reads in promoters of 0.46 and high reproducibility ($R > 0.99$) between biological duplicates (**Supplementary Fig. 7a-b; Supplementary Tables 1**). After data aggregation, 150,593 peaks were detected, resulting in a highly dimensional and sparse scATAC-seq matrix (4.2% of non-zeros).

Next, we performed data integration for batch effect removal using Harmony[31]. For comparison, we used a dimension reduced matrix from either LSI (Cusanovich2018), cisTopic, SnapATAC or scOpen. We annotated the scATAC-seq profiles using single nuclei RNA-seq (snRNA-seq) data of the same kidney fibrosis model from an independent study[32] via label transfer[33] to serve as cell labels. We then evaluated the batch correction results using silhouette score and clustering. Notably, we observed that clusters based on scOpen were more similar to the transferred labels (higher ARI) than clusters based on competing methods (**Fig. 3a**). Furthermore, scOpen also provided better distance metrics and visualisation than competing methods (**Supplementary Fig. 7c-e; Supplementary Fig. 8**). These results support the discriminative power of scOpen in this large and complex dataset.

Next, we annotated the clusters of scOpen by using known marker genes and transferred labels after removing doublets with ArchR[34]. We identified all major kidney cell types including proximal tubular cells, distal/connecting tubular cells, collecting duct and loop of Henle, endothelial cells, fibroblasts as well as the rare populations of podocytes and lymphocytes (**Fig. 3b**; **Supplementary Fig. 9a**). Lymphocytes were not described in the previously scRNA-seq study[32], which supports that importance of an annotation of scATAC-seq clusters independently of scRNA-seq label transfer. Of particular interests were cell types with population changes during progression

of fibrosis (**Fig. 3c**; **Supplementary Fig. 9b-d**). We observed an overall decrease of normal proximal tubular, glomerular and endothelial cells and increase of immune cells as expected in this fibrosis model with tubule injury, influx of inflammatory cells and capillary loss[35,36]. Importantly, we detected an increased PT sub-population, which we characterised as injured PT by an increased accessibility around the PT injury markers Vcam1 and Kim1(Havrc1)[37](**Supplementary Fig. 9a**).

Dissecting cell specific regulatory changes in fibrosis

Next, we adapted HINT-ATAC[4] to dissect regulatory changes in scATAC-seq clusters. For each cluster, we created a pseudo-bulk ATAC-seq library by combining reads from single cells in the cluster. We then performed footprinting analysis and estimated TF activity scores for all footprint supported motifs. We only kept TFs with changes (high variance) in TF activity scores among clusters. We focused here on clusters associated to proximal tubular cells (PT), fibroblasts and immune cells, as these represent key players in kidney remodelling and fibrosis after injury. As shown in **Fig. 3d**, the TF activity scores capture regulatory programs associated with these 3 major cell populations. Interestingly, injured PTs have overall lower TF activity scores of all TFs of the PT cluster. TFs with high decrease in activity in injured PTs include Rxra, which is important for the regulation of calcium homeostasis in tubular cells[38], and Hnf4a, which is important in proximal tubular development[39] (**Fig. 3e**). Footprint profiles of Rxra and Hnf4a in injured PTs display a gradual loss of TF activity over time indicating that injured PT acquire a de-differentiated phenotype during fibrosis progression and tubular dilatation (**Fig. 3f**). A group of TFs with high activity scores in injured PTs also have increased TF activity scores in fibroblasts (Smad2:Smad3 and Batf:Jun) indicating shared regulatory programs in these cells. Smad proteins are downstream mediators of TGFβ signalling, which is a known key player of fibroblast to myofibroblast differentiation and fibrosis [40]. The high activity of Smad2::Smad3 also indicate a role of TGFβ in the de-differentiation of injured PTs. Interestingly, Smad2:Smad3 reach a peak in TF activity level at day 2 after UUO in injured PTs (**Fig. 3f**), which indicate these TFs are activated post-transcriptionally.

scOpen reveals transcription factors driving myofibroblast differentiation

A key process in kidney injury is fibrosis, which is caused by the differentiation of fibroblasts and pericytes to matrix secreting myofibroblasts[41]. To dissect potential differentiation trajectories, we performed a diffusion map embedding of the fibroblasts (**Fig. 4a**), which revealed the presence of three major branches formed by fibroblasts, pericytes and myofibroblasts, as supported by the expression of Scara5, Ng2(Cspg4), Postn and Col1a1 (**Supplementary Fig. 10**)[41,42].

We next created a cellular trajectory across the differentiation from fibroblasts to myofibroblasts using ArchR(**Fig. 4a**; **Supplementary Fig. 10c**). We observed that there is an increase in cells after injury (Day 2 and Day 10) along the trajectory (**Fig. 4b**). We next characterised TFs by correlating their gene activity with TF activity along the trajectory (**Fig. 4c**) and ranked these by their correlation (**Supplementary Fig. 10d**). The correlation of Runx1,

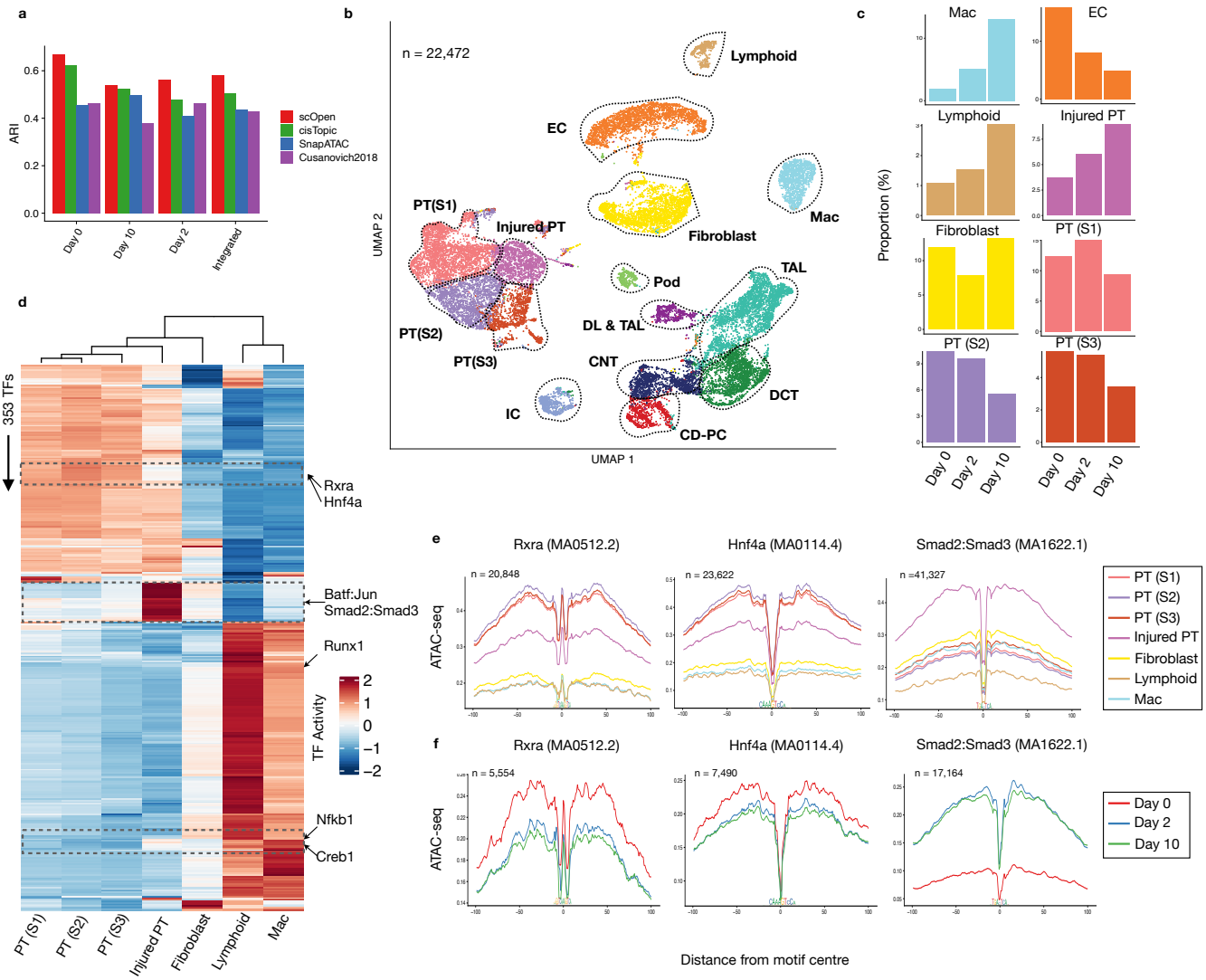

Fig. 3. scOpen characterises progression of kidney fibrosis. a, ARI values (y-axis) contrasting clustering results and transferred labels using distinct dimensional reduction methods for scATAC-seq. Clustering was performed by only considering UUO kidney cells on day 0 (WT), day 2 or day 10 or the integrated data set (all days). **b,** UMAP of the integrated UUO scATAC-seq after doublet removal with major kidney cell types: fibroblasts, descending loop of Henle and thin ascending loop of Henle (DL & TAL); macrophages (Mac), Lymphoid (T and B cells), endothelial cells (EC), thick ascending loop of Henle (TAL), distal convoluted tubule (DCT), collecting duct-principal cells (CD-PC), intercalated cells (IC), podocytes (Pod) and proximal tubule cells (PT S1; PT S2; PT S3; Injured PT). **c,** Proportion of cells of selected clusters on either day 0, day 2 or day 10 experiments. **d,** Heatmap with TF activity score (z-transformed) for TFs (y-axis) and selected clusters (x-axis). We highlight TFs with decrease in activity scores in injured PTs (Rxra and Hnf4a), with high TF activity scores in injured PTs (Batf:Jun; Smad2:Smad3) and immune cells (Creb1; Nfkb1). **e,** Transcription factor footprints (average ATAC-seq around predicted binding sites) of Rxra, Hnf4a and Smad2::Smad3 for selected cell types. Logo of underlying sequences is shown below and number of binding sites is shown top-left corner. **f,** Transcription factor footprints of Rxra, Hnf4a and Smad2::Smad3 for injured PT cells in day 0, day 2 and day 10.

which has a well known function in blood cells[43], stood out, besides showing a steady increase in activity in myofibroblasts. Another TF with high correlation and similar myofibroblast specific activity was Twist2, which has a known role in epithelial to mesenchymal transition in kidney fibrosis[44] (**Fig. 4d**).

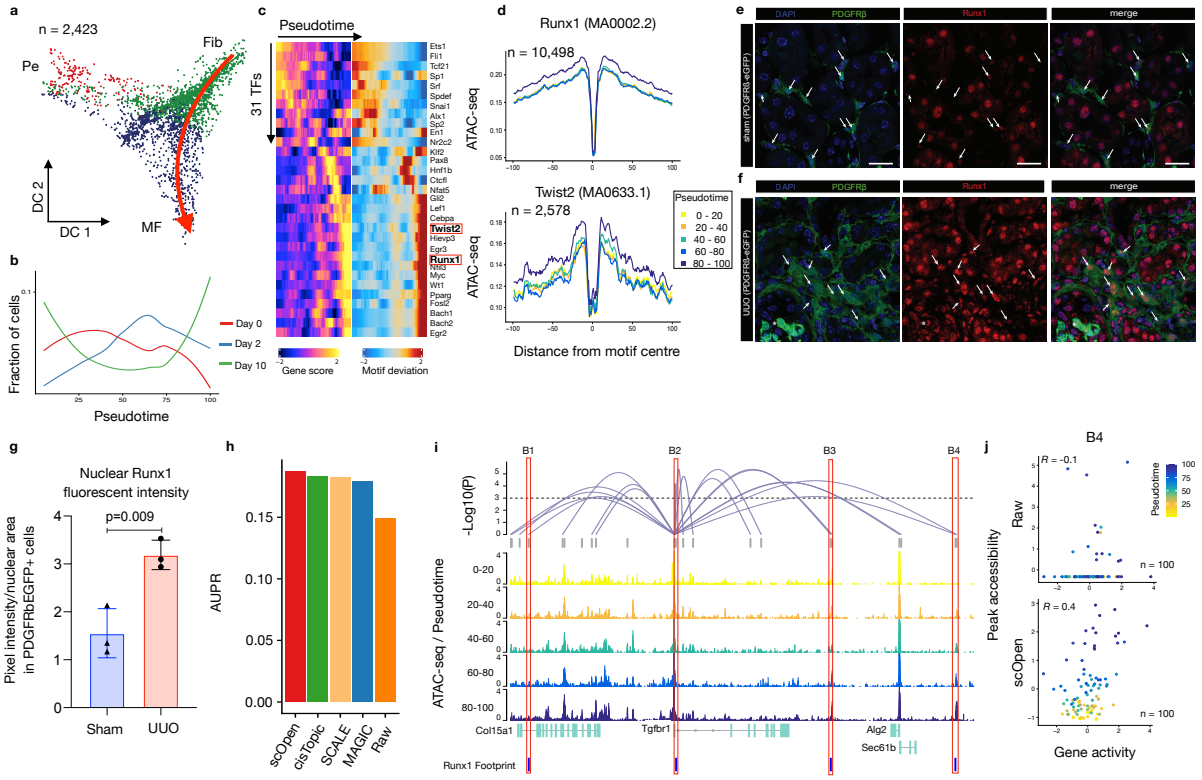

Fig. 4. Role of Runx1 in myofibroblast differentiation. a, Diffusion map showing sub-clustering of fibroblasts. Colours refer to sub-cell-types and arrow represents differentiation trajectory from fibroblast to myofibroblast. Pe (pericyte), Fib (fibroblast), MF (myofibroblast). **b,** Line plots showing cell proportion from day after UUO along the trajectory. **c,** Pseudotime heatmap showing gene activity (left) and TF motif activity (right) along the trajectory. **d,** Footprinting profiles of Runx1 and Twist2 binding sites along the trajectory. **e,** Immuno-fluorescence (IF) staining of Runx1 (red) in PDGFRb-eGFP mouse kidney. In sham operated mice Runx1 staining shows a reduced intensity in PDGFRb-eGFP+ cells compared to remaining kidney cells (arrows). **f,** Immuno-fluorescence (IF) staining of Runx1 (red) in PDGFRb-eGFP mouse kidney at 10 days after UUO as compared to sham. Arrows indicate Runx1 staining in expanding PDGFRb-eGFP+ myofibroblasts. **g,** Quantification of Runx1 nuclear intensity in PDGFRb-eGFP+ cells in sham vs. UUO mice (n=3). **h,** Performance of top performing imputation methods on the prediction of Runx1 target genes measured with AUPR. **i,** Peak-to-Gene links (top) predicted on scOpen matrix and associated to Tgfbr1 in fibroblast cells. Height of links represent it's significance. Dash line represents threshold of significance (FDR = 0.001). ATAC-seq tracks (below) were generated from pseudo-bulk profiles of fibroblast/myofibroblast cells with increasing pseudo time (0-20, 20-40, 40-60, 60-80 and 80-100). Binding sites of Runx1 (B1-B4) supported by ATAC-seq footprints and overlapping to peaks are highlighted on bottom. **j,** Scatter plot showing gene activity of Tgfbr1 and normalised peak accessibility from raw (upper) or scOpen imputed matrix (lower) for peak-to-gene link B4. Each dot represents cells in a given pseudotime and the overall correlation is shown on left-upper corner.

To validate the yet uncharacterized role of Runx1 in myofibroblasts, we performed immunostaining and quantification of Runx1 signal intensity in transgenic PDGFRb-eGFP mice that genetically tag fibroblasts and myofibroblasts[41,45]. Runx1 staining in control mice (sham) revealed positive nuclei in tubular epithelial cells and rarely in PDGFRbeGFP+ mesenchymal cells (**Fig. 4e**). In kidney fibrosis after UUO surgery (day 10), Runx1 staining intensity increased significantly in PDGFRb+ myofibroblasts (**Fig. 4f-g**). Next, we performed lentiviral overexpression experiments and RNA-sequencing in a human kidney PDGFRb+ fibroblast cell-line that we have generated[41] to ask whether Runx1 might be functionally involved in myofibroblast differentiation in humans (**Supplementary**

Fig. 11a-b). Runx1 over-expression led to reduced proliferation (**Supplementary Fig. 11c**) and strong gene expression changes (**Supplementary Fig. 11d**). GO and pathway enrichment analysis indicated enrichment of cell adhesion, cell differentiation and TGFB signalling following Runx1 overexpression (**Supplementary Fig. 11e**). Various extracellular matrix genes (Fn1, Col13A1) as well as a TFGB receptor (Tgfbr1) and Twist2 were up-regulated following Runx1 overexpression (**Supplementary Fig. 11d; Supplementary File 1**). Furthermore, we observed increased expression of the myofibroblast marker gene Postn after Runx1 overexpression. Altogether, this suggests that Runx1 might directly drive myofibroblast differentiation of human kidney fibroblasts since overexpression reduced cell-proliferation an induced expression of various myofibroblast genes.

Identification of Runx1 target genes

Another important application of scATAC-seq is the prediction of *cis*-regulatory DNA-interactions (peak-to-gene links) by measuring the correlation between gene activity and reads counts in proximal peaks. To compare the impact of imputation on this task, we predicted peak-to-gene links in fibroblasts on distinct scATAC-seq matrices using ArchR[34] after imputation wiht top performing imputation methods. The use of imputation methods led to improved signals on peak-to-gene links predictions as indicated by higher correlation values after imputation (**Supplementary Fig. 12a-b**). We considered all genes with at least one link, where the peak has a footprint supported Runx1 binding site, as Runx1 targets. We then compared the predicted Runx1 targets from distinct scATAC-seq imputed matrices with differential expressed genes after Runx1 overexpression (true labels). Interestingly, all imputation methods obtained higher AUPR values than the use of a raw matrix, while scOpen obtained the highest AUPR (**Fig. 4h; Supplementary Fig. 12c**). Among others, scOpen predicted Tgfbr1 and Twist2 as prominent Runx1 target genes (**Fig. 4i; Supplementary Fig. 12d**). We observed several peaks with high peak-to-gene correlation, increasing accessibility upon myofibroblast differentiation and presence of Runx1 binding sites. The positive impact of imputation was clear when observing scatter plots contrasting gene activity and peak accessibility of these peak-to-gene links (**Fig. 4j; Supplementary Fig. 12e-i**). These results suggest that Runx1 is an important regulator of myofibroblast differentiation by regulating the EMT related TF Twist2 and by amplifying TGFB signalling by increasing the expression of a TGFB receptor 1 and affecting expression of extracellular matrix genes. Altogether, these results uncover a complex cascade of regulatory events across cells during progression of fibrosis and reveal an yet unknown function of Runx1 in myofibroblast differentiation in kidney fibrosis.

Discussion

In ATAC-seq, Tn5 generates a maximum of 2 fragments per cell in a small ($\sim$200bp) open chromatin region. Subsequent steps of the ATAC-seq protocol cause loss of a large proportion of these fragments. For example, only DNA fragments with the two distinct Tn5 adapters, which are only present in 50% of the fragments, are amplified in the PCR step[46]. Further DNA material losses occur during single cell isolation, liquid handling, sequencing or

by simple financial restrictions of sequencing depth. Assuming that 25% of accessible DNA can be successfully sequenced, we expect that 56%[1] of accessible chromatin sites will not have a single digestion event causing the so-called dropout events. Despite this major signal loss, imputation and denoising has been widely ignored in the scATAC-seq literature[5,6,8,9,12,13] and common scATAC-seq pipelines ArchR[34].

We demonstrated here that scOpen estimated matrices have a higher recovery of dropout events and also improved distance and clustering results, when compared to imputation methods for scRNA-seq[14,22–25] and the few available imputation methods tailored for scATAC-seq (cisTopic-impute[10], SCALE[17]). scOpen also presented very good scalability with lowest memory requirements and tractable computational time on large data sets. From a methodological perspective, scOpen is the only method performing regularisation of estimated models to prevent over-fitting. This is in line with a previous study, which indicated over-fitting as one of the largest issues on scRNA-seq imputation[47]. Moreover, it is also possible to use the scOpen factorised matrix as a dimension reduction. We have shown that both dimension reduced and imputed matrices from scOpen scOpen displayed the best performance on distance representation and clustering when compared to diverse state-of-art scATAC-seq dimension reduction/clustering pipelines (cisTopic, SnapATAC and Cusanovich et. al 2018). [2]

Finally, we have demonstrated that the use of scOpen corrected matrices improves the accuracy of existing state-of-art scATAC-seq methods (cisTopic[10], chromVAR[12], Cicero[13]). Particularly positive results were obtained in prediction of chromatin conformation with Cicero, where all methods perform better than raw-matrices. Cicero works by measuring correlation between pairs of proximal links. Due to the fact that dropout events are independent for two regions, it is not surprising that imputation has strong benefits. This is equivalent to observations from van Dijk et al., 2018[14] in the context of scRNA-seq, where the prediction of gene-gene interactions after MAGIC imputation were significantly improved. Altogether, these results support the importance of dropout event correction with scOpen in any computational analysis of scATAC-seq. Of note, a sparsity similar to scATAC-seq are also expected in single cell protocols based on DNA enrichment such as scChIP-seq[48,49], scCUT&Tag[50] or scBisulfite-seq[51]. Denoising and imputation of count matrices from these protocols represents a future challenge.

Moreover, we used scOpen to characterise complex cascades of regulatory changes associated to kidney injury and fibrosis. Our analyses demonstrate that major expanding population of cells, i.e. injured PTs, myofibroblasts and immune cells, share regulatory programs, which are associated with cell de-/differentiation and proliferation. Of all methods evaluated, scOpen obtained best clustering results in the kidney cell repertoire using a scRNA-seq on the same kidney injury model as a reference. Trajectory analysis identified Runx1 as the major TF driving myofibroblast differentiation, which was validated by Runx1 staining in mouse model and by lentiviral overexpression studies in human PDGFRb+ kidney cells. Computational prediction with peak-to-gene links combined with footprint supported Runx1 binding sites indicates the role of Runx1 in regulation of Tgfbr1 and Twist2. These were validated

[1] We assume digestion events follow a binomial distribution.

[2] Of note the new ArchR pipeline is equivalent to Cusanovich et al. 2018 and based on the same dimension reduction/clustering methods (LSI).

on over-expression experiments in human fibroblasts. Altogether, results suggests that Runx1 makes fibroblasts more sensitive to TGFB signalling via increasing expression the TGFB receptors. Runx1 has recently been reported as a potential inducer of EMT in proximal tubular cells [52] while a role in renal myofibroblasts has not been shown. The role of Runx1 as driver of scar formation was recently described in the zebrafish heart[53]. After injury Runx1 was up-regulated in endocardial cells and thrombocytes that expressed collagens shown by single-cell sequencing. Runx1 deficiency caused reduced myofibroblast formation and enhanced recovery. To this end, inhibiting Runx1 could lead to reduced myofibroblast differentiation and increased endogenous repair after fibrogenic organ injuries in the kidney and heart. Our results shed novel light into mechanisms of myofibroblasts differentiation driving kidney fibrosis and chronic kidney disease (CKD). Altogether, this demonstrates how scOpen can be used to dissect complex regulatory process by footprinting analysis combined with peak-to-gene link predictions.

Methods

scOpen

scOpen aims to simultaneously impute and reduce the dimension of a scATAC-seq matrix. Let $X \in R^{m \times n}$ be the scATAC-seq matrix, where X_{ij} is the number of cutting sites in peak i and cell j; m is the total number of peaks and n is the number of cells. We first define a binary open/closed chromatin matrix $\hat{X} \in \{0,1\}^{m \times n}$, i.e.

$$X_{ij}^{'} = \begin{cases} 1 & X_{ij} > 0 \\ 0 & X_{ij} = 0 \end{cases} \tag{1}$$

where 1 indicates the peak i is open and 0 indicates closed in cell j. Next, we calculate a score for peak i and cell j by applying term frequency–inverse document frequency (TF-IDF) transformation[54]

$$X_{ij}^{''} = \frac{X_{ij}^{'}}{\sum_{p=1}^{m} X_{pj}^{'}} \cdot log(\frac{n}{\sum_{q=1}^{n} X_{iq}^{'}}) \tag{2}$$

This score represents how important the peak i is for cell j. Next, we normalise the TF-IDF matrix as

$$V_{ij} = \frac{X_{ij}^{''}}{\sqrt{\sum_{k=1}^{m} X_{kj}^{''\,2}}} \tag{3}$$

We next impute the matrix by minimisation of the following optimisation problem:

$$\hat{M} = argmin \sum_{i,j} (M_{ij} - V_{ij})^2 + \lambda ||M||_*, \quad s.t. \quad M_{ij} \geq 0 \tag{4}$$

where $||M||_* = \sum_i^k \sigma_i(M)$ is the nuclear norm of matrix M, and σ_i denotes the ith largest singular value of M. The first item is the estimator of square loss for each element in M and λ is the regularisation parameter, which aims to prevent the model from over-fitting and set to 1 as default value. To solve this problem, we assume that M is a low-rank matrix with rank k and it can be written as:

$$\min_{W,H} f(W,H) = \sum_{ij} ((WH)_{ij} - V_{ij})^2 + \frac{\lambda}{2}||W||^2 + \frac{\lambda}{2}||H||^2, \quad s.t. \quad (WH)_{ij} \geq 0 \tag{5}$$

where $W \in \mathbb{R}^{m \times k}, H \in \mathbb{R}^{k \times n}$. This constrained optimisation problem is solved by using cyclic coordinate descent (CCD) methods[55]. This method iteratively updates the variable w_{it} in W to z by solving the following one-variable sub-problem.

$$\min_z f(z) = \sum_{j=1}^{n} ((\sum_{t' \in k} w_{it'} h_{t'j} - w_{it} h_{tj}) + z h_{tj} - \frac{1}{1-\rho_j} \hat{X}_{ij})^2 + \frac{\lambda}{2} z^2, \quad s.t. \quad z \geq 0 \tag{6}$$

Likewise, the elements in H can be updated with similar update rule. The above iteration is carried out until a termination criterion is met, e.g. number of iteration performed. Afterwards, we calculate M as the product of W and H to obtain the scOpen imputed matrix or consider H as scOpen reduced matrix. This algorithm has a theoretical time complexity of $O((m+n)k)$ for a single iteration and thus is scalable for large datasets.

Selection of hyper-parameters in scOpen

There are two hyper-parameters in scOpen, i.e., rank of the matrix k and regularisation parameter λ. Rank k determines the intrinsic dimensions of a matrix and thus is highly dataset-specific. To select an appropriate value of k, we first input a number of ranks and generated an residual sum of squares (RSS) curve (Eq. 4) in a pre-defined interval (2-30 as default). Next, we use a knee point detection method [19], which finds a k with best trade-off between fit error and model complexity. We make use of a simulated scATAC-seq dataset as described below to evaluate the impact of λ and k in either the imputation and clustering performance (see **Supplementary Fig. 1**). We use optimal settings ($lambda = 1$ and knee detection from an interval of 2-30 in further results.

scATAC-seq simulation dataset

To generate a simulation scATAC-seq dataset, we downloaded bulk ATAC-seq data of 13 FACS-sorted human primary blood cell types from gene expression omnibus (GEO) with accession number GSE74912[56]. For each cell type, we processed the data similarly as in[12]. First, the downloaded files were converted to FastQ using SRA toolkit (`http://ncbi.github.io/sra-tools/`). Next, adapter sequences and low-quality ends were trimmed from FastQ files using Trim Galore[57]. Reads were mapped to the genome hg19 using Bowtie2[58] with the following parameters ($-$X 2000 $--$very$-$sensitive $--$no$-$discordant), allowing paired end reads of up to 2 kb to align. Then, reads mapped to chrY, mitochondria and unassembled "random" contigs were removed. Duplicates were also removed with Picard[59] and reads were further filtered for alignment quality of >Q30 and required to be properly paired using samtools[60]. Peaks were called using MACS2[61] with the following parameters ($--$keep$-$dup auto $--$call$-$summits). We next merged the peaks from all cell types to create a unique peaks list. We then created a peak cell-type matrix by offsetting $+$4 bp for forward strand and $-$5bp for reverse strand to represent the cleavage event centre[1,4] and counting the number of read start sites per cell type in each peak. This provides a cell type vs peak matrix A, where a_{ij} indicates the number of reads for cell j in peak i.

We next used this bulk ATAC-seq counts matrix A to simulate a scATAC-seq counts matrix X by improving the simulation strategy proposed in[20]. Specifically, given m peaks and T cell types, we define the accessibility $x_{ij} \in \{0,1\}$

of a single cell j from the cell type t in peak i as:

$$
\begin{aligned}
x_{ij} &\sim Bernoulli(p_i^t) \\
p_i^t &= r_i^t \cdot n_j \cdot (1-q) + (\frac{1}{m}) \cdot n_j \cdot q \\
r_i^t &= \frac{a_{it}}{\sum_{k=1}^{m} a_{kt}} \\
n_j &= N_j \cdot f \\
N_j &\sim NB(r, p)
\end{aligned}
\tag{7}
$$

where p_i^t denotes probability of cell type t being accessible in peak i, q is a noise parameter, n_j denotes the number of reads in peaks for single cell j, f denotes the fraction of reads in peaks (FRiP) and N_j denotes the total number of reads for cell j. N_j is sampled from a negative binomial distribution, whose parameters were estimated from a real scATAC-seq dataset. We simulated 200 cells per cell type using above process and used noise $q = 0.6$ and FRiP $f = 0.3$. Our approach differs from[20] by sampling the number of reads per cell from a negative binomial distribution rather than using a fixed number and the introduction of the FRiP parameters.

scATAC-seq benchmarking datasets

The cell line dataset was obtained by combining single cell ATAC-seq data of BJ, H1-ESC, K562, GM12878, TF1 and HL-60 from [5], which was downloaded from GEO with accession number GSE65360. The hematopoiesis dataset includes scATAC-seq experiments of sorted progenitor cells populations: hematopoietic stem cells (HSC), multipotent progenitors (MPP), lymphoid-primed multi-potential progenitors (LMPP), common myeloid progenitors (CMP), common lymphoid progenitors (CLP), granulocyte-macrophage progenitors (GMP), megakaryocyte–erythroid progenitors (MEP) and plasmacytoid dendritic cells (pDC)[6]. Sequencing libraries were obtained from GEO with accession number GSE96769. In both datasets, the original cell types were used as true labels for clustering as in previous work[9,10]. The T cell dataset is based on human Jurkat T cells, memory T cells, naive T cells and Th17 T cells obtained from GSE107816[8]. Labels of memory, naive and Th17 T cells were provided in Satpathy et al.[8] by comparing scATAC-seq profiles with bulk ATAC-seq of corresponding T cell subpopulations. For each of these three datasets, we pre-processed the data per cell as described above and only kept cells with at least 500 unique fragments. We then created a pseudo-bulk ATAC-seq library by merging the obtained scATAC-seq profiles and called peaks using MACS2[61]. The peaks were extended $\pm$250bp from the summits as in[1] and peaks overlapping with ENCODE blacklists (`http://mitra.stanford.edu/kundaje/akundaje/release/blacklists/hg19-human/`) were removed. we next constructed a peak by cell counts matrix. To the test scalability of imputation methods, we also included a multiome PBMC dataset with 10,000 cells (`https://support.10xgenomics.com/single-cell-multiome-atac-gex/datasets`). This dataset was generated using the Chromium Single Cell Multiome ATAC + Gene Expression assay. We use here the cell types as

annotated by the 10X Genomics R&D team using only the scRNA-seq data. See Supplementary Table 1 for complete statistics associated to these data sets.

Comparison between scOpen and competing imputation methods

We compared the performance of scOpen with 8 competing imputation approaches, i.e., MAGIC[14], SAVER[22], scImpute[23], DCA[24], cisTopic[10], scBFA[25], SCALE[17] and PCA. We performed imputation with these algorithms (see details below) on the benchmarking datasets.

We first tested if the imputed matrix can recover the true signal for each single cell. To this end, we used cell labels from each dataset to aggregate the ATAC-seq profiles and performed peak calling to find cell specific OC regions. OC regions present in a particular cell type were considered as trues and OC regions not present in that cell type as negatives. For a particular cell, we can the obtain true positives and true negatives by comparing the labels of the corresponding cell type with the presence of reads (or openness score) in that OC region and single cell. We use these statistics to measure the area under the precision recall curve (AUPR)[62] for each cell.

Next, we evaluated the imputed matrix using mean silhouette score of cells[63]. For a given cell x:

$$silhouette(x) = \frac{b(x) - a(x)}{max(a(x), b(x))} \tag{8}$$

where $a(s)$ is the average distance between x and the other cells of the same class, and $b(x)$ is the average distance between x and cells in the closest different class. The distance was calculated as 1 - Pearson correlation. A higher silhouette score indicates higher similarity of a cell to cells of the same cell type than cells from other cell types.

We next tested if the imputed matrix improves cell clustering. We applied PCA (50 PCs) for each of the imputed matrix and clustered cell using k-medoids and hierarchical clustering methods with 1 - Pearson correlation as distance. We also used t-SNE[64] embedding as input and euclidean as distance. We also tested different number of clusters, e.g. k and k + 1, where k is the true number of clusters. We used adjusted Rand index (ARI) to evaluate the clustering results[27] with labels from benchmarking data sets (See **Supplementary Fig. 1** for experimental design). The adjusted Rand index measures similarity between two data clustering by correcting the chance of grouping elements. Specifically, given two partitions of a dataset D with n cells, $U = \{U_1, U_2, \cdots U_r\}$ and $V = \{V_1, V_2, \cdots, V_s\}$, the number of common cells for each cluster i and j can be written as:

$$c_{ij} = |U_i \cap V_j| \tag{9}$$

where $i \in \{1, 2, \cdots, r\}$ and $j \in \{1, 2, \cdots, s\}$. The ARI can be calculated as follows:

$$ARI = \frac{\sum_{ij}\binom{c_{ij}}{2} - [\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}]/\binom{n}{2}}{\frac{1}{2}[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}] - [\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}]/\binom{n}{2}} \tag{10}$$

where $a_i = \sum_{j=1}^{s} c_{ij}$ and $b_j = \sum_{i}^{r} c_{ij}$, respectively. The ARI has a maximum value 1 and an expected value 0, with 1 indicating that the data clustering are the exactly same and 0 indicating that the two data clustering agree randomly.

MAGIC

MAGIC is an algorithm for alleviating sparsity and noise of single cell data using diffusion geometry[14]. We downloaded MAGIC from https://github.com/KrishnaswamyLab/MAGIC and applied it on the count matrix with default setting. Prior to MAGIC, the input was normalised by library size and root squared, as suggested by the authors[14].

SAVER

SAVER is a method that recovers the true expression level of each gene in each cell by borrowing information across genes and cells[22]. We obtained SAVER from https://github.com/mohuangx/SAVER and ran it on the normalised tag count matrix with the default parameters.

scImpute

scImpute is a statistical method to accurately and robustly impute the dropouts in scRNA-seq data[23]. We downloaded scImpute from https://github.com/Vivianstats/scImpute and executed it using the default setting except for the number of cell clusters which is used to determine the candidate neighbours of each cell by scImpute. We defined this as the true cluster number for each benchmarking dataset.

DCA

DCA is a deep auto-encoder network for denoising scRNA-seq data by taking the count structure, over-dispersed nature and sparsity of the data into account[24]. We obtained DCA from https://github.com/theislab/dca and ran it with default setting.

cisTopic-impute

cisTopic is a probabilistic model to simultaneously identify cell states (topic-cell distribution) and cis-regulatory topics (region-topic distribution) from single cell epigenomics data[10]. We downloaded it from https://github.com/aertslab/cisTopic and ran it with different numbers of topics (from 5 to 50). The optimal number of topics was selected based on the highest log-likelihood as suggested in[10]. We then multiplied the topic-cell and the region-topic distributions to obtain the predictive distribution[10], which describes the probability of each region in each cell and is used as imputed matrix for clustering and visualisation. We call this method as `cisTopic-impute`.

scBFA

scBFA is a detection-based model to remove technical variation for both scRNA-seq and scATAC-seq by analysing feature detection patterns alone and ignoring feature quantification measurements[25]. We obtained scBFA from

https://github.com/quon-titative-biology/scBFA and ran it on the raw count matrix using default parameters.

SCALE

SCALE combines the variational auto-encoder (VAE) and the Gaussian Mixture Model (GMM) to model the distribution of high-dimensional sparse scATAC-seq data[17]. We downloaded it from https://github.com/jsxlei/SCALE and ran it with default setting. We used option *–impute* to get the imputed data.

PCA

We also included principal component methods (termed here as PCA) on incomplete data sets as a control for comparison. We installed R package missMDA[26] and performed imputation with function imputePCA with default settings.

Evaluation of time and memory requirement of imputation methods

To compare the memory and running time requirement of each imputation method, we ran all of them on a dedicated HPC node with the same computation resources quota, i.e., 180GB memory, 120 hours time and 4 CPUs. For DCA and SCALE, two deep learning-based methods, we used GPU with 16GB memory. We measured the max memory usage during the running of a method and observed that all methods but PCA, SAVER and DCA can successfully generate the imputed matrix for all datasets (**Fig. 1b**). For multi-omics PBMC dataset, DCA failed due to GPU memory issue and we could not obtain results from PCA and SAVER after 120 hours of running.

Comparison between scOpen and competing dimension reduction methods

We next compared the performance of scOpen with cisTopic[10], SnapATAC[?] and latent semantic indexing (LSI) (termed here as Cusanovich2018)[65] for dimension reduction of scATAC-seq data. We applied these methods to obtain a low-dimension matrix from each dataset (detailed below) and measured the mean silhouette score[63] (See Fig. 1e).

cisTopic

We executed cisTopic as described above and used the topic-cell distribution as dimension reduced matrix.

SnapATAC

SnapATAC is a software package for analysing scATAC-seq datasets[11]. Instead of using peak annotation as features, it resolves cellular heterogeneity by directly comparing the similarity in genome-wide accessibility profiles between cells. Furthermore, SnapATAC uses Nyström method to generate a low rank embedding for large-scale dataset which enables the analysis of scATAC-seq up to a million cells. We installed SnapATAC from https://github.com/r3fang/SnapATAC and followed the tutorial from https://github.com/r3fang/SnapATAC/blob/master/examples/10X_brain_5k/README.md to perform dimension reduction for benchmarking datasets.

Cusanovich2018

Cusanovich2018 first segments the genome into 5kb windows and then scored each cell for any insertions in these windows, generating a large, sparse, binary matrix of 5kb windows by cells. Based on this matrix, the top 20,000 most commonly used sites were retained. Then, the matrix was normalised and re-scaled using the term frequency-inverse document frequency (TF-IDF) transformation. Next, singular value decomposition (SVD) was performed to generate a PCs-by-cells low dimension matrix.

Benchmarking of scATAC-seq downstream analysis methods

Next, we compared the performance of state-of-art scATAC-seq methods (scABC, chromVAR and Cicero) when presented with either scOpen imputed or raw scATAC-seq matrix. The rationale is if we improve the count matrix by imputation, we should be able to improve downstream analysis. Note that scABC is the only method providing a clustering solution. chromVAR and Cicero transform the scATAC-seq matrices into transcription factor and gene space. We here again evaluated the results based on clustering accuracy with methods used as standard by these pipelines, i.e., hierarchical clustering with complete agglomeration method for chromVAR and k-medoids for Cicero. Moreover, we evaluated the co-accessible links predicted by Cicero between using scOpen imputed or raw counts matrix.

scABC

scABC is an unsupervised clustering algorithm for single cell epigenetic data[9]. We downloaded it from https://github.com/SUwonglab/scABC and executed according to the tutorial https://github.com/SUwonglab/scABC/blob/master/vignettes/ClusteringWithCountsMatrix.html.

chromVAR

chromVAR is an R package for analysing sparse chromatin-accessibility data by measuring the gain or loss of chromatin accessibility within sets of genomic features, as regions with sequence predicted transcription factor (TF) binding sites[12]. We obtained chromVAR from https://github.com/GreenleafLab/chromVAR and executed to find gain/loss of chromatin accessibility in regions with binding sites of 571 TF motifs obtained in JASPAR version 2018[66].

Cicero

Cicero is a method that predicts co-accessible pairs of DNA elements using single-cell chromatin accessibility data[13]. Moreover, Cicero provides a gene activity score for each cell and gene by assessing the overall accessibility of a promoter and its associated distal sites. This matrix was used for clustering and visualisation of scATAC-seq. We obtained Cicero from https://github.com/cole-trapnell-lab/cicero-release and executed it according to the document provided by https://cole-trapnell-lab.github.io/cicero-release/docs/.

Chromosomal conformation experiments with Cicero

We used conformation data as true labels to evaluate co-accessible pairs of cis-regulatory DNA as detected by Cicero on GM12878 cells. We obtained scATAC-seq matrix of GM12878 cells from GEO (GSM2970932). For evaluation, we downloaded promoter-capture (PC) Hi-C data of GM12878 from GEO (GSE81503), which use CHiCAGO[67] score as physical proximity indicator. We also downloaded ChIA-PET data of GM12878 from GEO (GSM1872887), which used the frequency of each interaction PET cluster to represent how strong an interaction is. We considered all obtained links, as provided by these data sets, as true interactions as in[13]. Next, we replicated the evaluation analysis performed in Fig. 4 of ref.[13] and contrasted the results of Cicero with raw or matrices obtained after scOpen imputation. Next, we use the built-in function *compare_connections* of Cicero to define the true labels for predicted co-accessibility links. Using the correlation as prediction, we finally computed the the area of precision and recall curve (AUPR) with *pr.curve* function from R package PRROC[68].

scATAC-seq UUO mouse kidney datasets

Animal experiments

Unilateral Ureter Obstruction (UUO) was performed as previously described[30]. Shortly, the left ureter was tied off at the level of the lower pole with two 7.0 ties (Ethicon) after flank incision. One C57BL/6 male mouse (age 8 weeks) was sacrificed on day 0 (sham), day 2 and 10 after the surgery. Kidneys were snap-frozen immediately after sacrifice. Pdgfrb-BAC-eGFP reporter mice (for staining experiments, age 6-10 weeks, C57BL/6) were developed by N. Heintz (The Rockefeller University) for the GENSAT project. Genotyping of all mice was performed by PCR. Mice were housed under specific pathogen–free conditions at the University Clinic Aachen. Pdgfrb-BAC-eGFP were sacrificed on day 10 after the surgery. All animal experiment protocols were approved by the LANUV-NRW, Düsseldorf, Germany. All animal experiments were carried out in accordance with their guidelines.

scATAC experiments

Nuclei isolation was performed as recommended by 10X Genomics (demonstrated protocol CG000169). The nuclei concentration was verified using stained nuclei in a Neubauer chamber with trypan-blue targeting a concentration of 10.000 nuclei. Tn5 incubation and library prep followed the 10X scATAC protocol. After quality check using Agilent BioAnalyzer, libraries were pooled and run on a NextSeq in 2x75bps paired end run using three runs of the the NextSeq 500/550 High Output Kit v2.5 Kit (Illumina). This results in more than 600 million reads.

UUO data pre-processing

We used Cell-Ranger ATAC (version-1.1.0) pipeline to perform low level data processing (`https://support.10xgenomics.com/single-cell-atac/software/pipelines/latest/algorithms/overview`). We first demultiplexed raw base call files using cellranger-atac mkfastq with its default setting to generate FASTQ files for each flowcell. Next, cellranger-atac count was applied to perform read trimming and filtering and alignment. We

then estimated Transcription Start Site (TSS) enrichment score using the obtained fragment files and filtered low quality cells using TSS score of 8 and number of unique fragments of 1,000 as thresholds. The obtained barcodes are considered as valid cells for following analysis.

UUO data dimension reduction, data integration and clustering

We next performed peak calling using MACS2 for each sample and merged the peaks to generate a union peaks set, which was used as features to create a peak by cell matrix. For comparison, we applied distinct methods, i.e., scOpen, cisTopic, SnapATAC and LSI/Cusanovich2018, to the matrix and used the dimension reduced matrix for data integration, clustering and visualisation. Next, we used Harmony[31] to integrate the scATAC-seq profiles from different conditions (day 0, day 2 and day 10) using either LSI/Cusanovich2018, cisTopic, scOpen or SnapATAC dimension reduced matrix as input. Specifically, we created an Seurat object for each of the low-dimension matrix and ran Harmony algorithm with the function *RunHarmony*. We then used k-medoids to cluster the cells taking batch-corrected low-dimension matrix as input. The number of clusters was set to 17 given that the single-nucleus RNA-seq that we used as reference for annotation identified 17 unique cell types (See below).

Label transfer and cluster annotation

To evaluate and annotate the clusters obtained from data integration, we downloaded a publicly available snRNA-seq dataset of the same fibrosis model (GSE119531) and performed label transfer using Seurat3[33]. This dataset contains 6147 single-nucleus transcriptomes with 17 unique cell types[32]. For label transfer, we used gene activity score matrix estimated by ArchR and transferred the cell types from snRNA-seq dataset to the integrated scATAC-seq dataset by using the function *FindTransferAnchors* and *TransferData* in Seurat3[33]. For benchmarking purposes, the predicted labels were used as true labels to compute ARI for evaluation of the clustering results and silhouette score for evaluation distances after using different dimension reduction methods as input for data integration (**Supplementary Fig. 7c-e**). We also performed the same analysis for each sample separately and evaluated the results (**Fig. 3a**).

For the biological interpretation, we estimated doublet scores using ArchR[34] and removed cells with doublet score > 2.5. Next, we named cluster by assigning the label with highest proportion of cells to the cluster and checking marker genes (**Supplementary Fig. 9a**). In total we recovered 16 unique cell types from the 17 labels, as two clusters (2 and 17) were annotated as TAL cells. Specifically, we denoted cluster 6, 1, 3 as proximal tubule (PT) S1, S2 and S3 cells. We annotated cluster 2 as thick ascending limb (TAL), cluster 5 as distal convoluted tubule (DCT), cluster 7 as collecting duct-principal cell (CD-PC), cluster 8 as endothelial cell (EC), cluster 9 as connecting tubule (CNT), cluster 10 as intercalated cell (IC), cluster 11 as fibroblast, cluster 12 as descending limb + thin ascending limb (DL TAL), cluster 13 as macrophage (MAC), cluster 16 as podocyte (Pod). Cluster 14 was identified as injured PT, which was not described in ref.[32], given the increased accessibility of marker Vcam1 and Havcr1 (**Supplementary Fig. 9a**). We also renamed the cells of cluster 15, which were label as Mac2 in ref[32], as lymphoid cells given that these cells express B and T cell markers Ltb and Cd1d, but not macrophage markers

C1qa and C1qb. Finally, cluster 4 was removed based on the doublet analysis.

Cell type specific footprinting with HINT-ATAC

We have adapted the footprinting based differential TF activity analysis from HINT-ATAC for scATAC-seq. In short, we created pseudo bulk atac-seq libraries by combining reads of cells for each cell type and performed footprinting with HINT-ATAC. Next, we predicted TF binding sites by motif analysis (FDR = 0.0001) inside footprint sequences using RGT (Version RGT-0.12.3; https://github.com/CostaLab/reg-gen). Motifs were obtained from JASPAR Version 2020[69]. We measured the average digestion profiles around all binding sites of a given TF for each pseudo bulk ATAC-seq library. We used then the protection score[4], which measures the cell specific activity of a factor by considering number of digestion events around the binding sites and depth of the footprint. Higher protection scores indicate higher activity (binding) of that factor. Finally, we only considered TFs with more than 1.000 binding sites and a variance in activity score higher than 0.3. See **Supplementary File 1** for complete activity scores results. We also performed smoothing for visualisation of average footprint profiles. In short, we performed a trimmed mean smoothing (5 bps window) and ignored cleavage values in the top 97.5% quantile for each average profile.

Identify trajectory from fibroblast to myofibroblast

We performed further sub-clustering of fibroblast cells on batch-corrected low-dimension scOpen matrix. In total, 3 clusters were obtained and annotated as pericyte (cluster 1), myofibroblast (cluster 2) and Scara5+ fibroblast (cluster 3) using known marker genes (**Supplementary Fig. 10a**), respectively. For visualisation, a diffusion map 2D embedding was generated using R package density[70]. Next, a trajectory from Scara5+ fibroblast to myofibroblast was created using function *addTrajectory* and visualised using function *plotTrajectory* from ArchR (**Supplementary Fig. 10c**).

Identify key TF drivers of myofibroblast differentiation

To identify TFs that drive this process, we first performed peak calling based on all fibroblasts using MACS2 to obtain specific peaks and then estimated motif deviation per cell using chromVAR. The deviation scores were normalised to allow for comparison between TFs. Next, we selected the TFs with high variance of deviation and gene activity score along the trajectory and calculated the correlation of TF activity and gene activity. This was done by using the function *correlateTrajectories* from ArchR. We only consider the 31 TFs with significant correlation (FDR < 0.1) (**Fig. 4c**). We then sorted the TFs by correlation, which identifies Runx1 as the most relevant TF for the differentiation (**Supplementary Fig. 10d**).

Prediction of Peak-to-Gene links

We obtained transcription start site (TSS) from annotation BSgenome.Mmusculus.UCSC.mm10 for each gene and extended it by 250k bps for both directions. Then, we overlapped the peaks from fibroblasts and the TSS regions using function *findOverlaps* to identify putative peak-to-gene links. We next created 100 pseudo-bulk ATAC-seq

profiles by assigning each cell to an interval along the trajectory of myofibroblast differentiation. The gene score matrix and peak matrix were aggregated according to the assignment to generate two pseudo-bulk data matrices. For each putative peak-to-gene link, we calculated the correlation between peak accessibility and gene activity. The p-values are computed using t distribution and corrected by Benjamini-Hochberg method. For comparison, we also performed matrix imputation using the four top methods, i.e., scOpen, SCALE, MAGIC and cisTopic, as evaluated by peaks recovering (**Supplementary Fig. 2b**) and computed the correlation based on imputed matrix.

Prediction and evaluation of Runx1 target genes

With each peak being associated to genes, we next sought to link Runx1 to its target genes. For this, we first performed footprinting using the peaks obtained from above and pseudo-bulk ATAC-seq profile to identify TF footprints. Next, we identified Runx1 binding sites using motif matching approach. We defined the genes that have at least one footprint-support binding site of Runx1 in their associated peaks as Runx1 target genes. We then used the peak-to-gene correlation as prediction between Runx1 and the target genes. This procedure was performed using the links estimated by different input data as described above, thus generating various prediction. To evaluate the results, we used the DE genes obtained from RNA-seq of Runx1 overexpression as true labels (See below), and computed the AUPR (**Fig. 4h**).

Immunofluorescence staining

Mouse kidney tissues were fixed in 4% formalin for 2 hours at RT and frozen in OCT after dehydration in 304%4 sucrose overnight. Using 5-10 μm cryosections, slides were blocked in 5% donkey serum followed by 1-hour incubation of the primary antibody, washing 3 times for 5 minutes in PBS and subsequent incubation of the secondary antibodies for 45 minutes. Following DAPI (4,6 – diamidino-2-phenylindole) staining (Roche, 1:10.000) the slides were mounted with ProLong Gold (Invitrogen, P10144). Cells were fixed with 3% paraformaldehyde followed by permeabilization with 0,3% TritonX. Cells were incubated with primary antibodies and secondary antibodies diluted in 2% bovine serum albumin in PBS for 60 or 30 minutes, respectively. The following antibodies were used: anti-Runx1 (HPA004176, 1:100, Sigma-Aldrich), AF647 donkey anti-rabbit (1:200, Jackson Immuno Research).

Confocal imaging and quantification

Images were acquired using a Nikon A1R confocal microscope using 40X and 60X objectives (Nikon). Raw imaging data was processed using Nikon Software or ImageJ. Systematic random sampling was applied to subsample of at least 3 representative areas per image of PDGFRbeGFP mice (n=3 mice per condition). Using QuPath nuclei were segmented and fluorescent intensity per nuclear size were measured of PDGFRbeGFP positive nuclei.

592 **Generation of a human PDGFRb+ cell line**

593 The cell line was generated using MACS separation (Miltenyi biotec, autoMACS Pro Separator,#130-092-545,
594 autoMACS Columns #130-021-101) of PDGFRb+ cells that were isolated from the healthy part of kidney cortex
595 after nephrectomy. The following antibodies were used for staining the cells and MACS procedure: PDGFRb (RD
596 #MAB1263 antibody, dilution 1:100) and anti-mouse IgG1-MicroBeads solution (Miltenyi, #130-047-102). The cells
597 were cultured in DMEM media (Thermo Fisher #31885) added 10% FCS and 1% penicillin/Streptomycin for 14
598 days. For immortalization (SV40-LT and HTERT) the retroviral particles were produced by transient transfection
599 of HEK293T cells using TransIT-LT (Mirus). Amphotropic particles were generated by co-transfection of plasmids
600 pBABE-puro-SV40-LT (Addgene #13970) or xlox-dNGFR-TERT (Addgene #69805) in combination with a packaging
601 plasmid pUMVC (Addgene #8449) and a pseudotyping plasmid pMD2.G (Addgene #12259) respectively. Using
602 Retro-X concentrator (Clontech) 48 hours post-transfection the particles were concentrated. For transduction the
603 target cells were incubated with serial dilutions of the retroviral supernatant (1:1 mix of concentrated particles
604 containing SV40-LT or rather hTERT) for 48 hours. At 72h after transfection the infected PDGFRb+ cells were
605 selected with 2 g/ml puromycin at 72h after transfection for 7 days.

606 ***Lentiviral overexpression of Runx1***

607 Runx1 vector construction and generation of stable Runx1-overexpressing cell lines. The human cDNA of Runx1 was
608 PCR amplified from 293T cells (ATCC, CRL-3216) using the primer sequences 5'- atgcgtatccccgtagatgcc-3' and 5'-
609 tcagtagggcctccacacgg-3'. Restriction sites and N-terminal 1xHA-Tag have been introduced into the PCR product us-
610 ing the primer 5'- cactcgaggccaccatgtacccatacgatgttccagattacgctcgtatccccgtagatgcc -3' and 5'- acggaattctcagtagggc-
611 ctccacac -3'. Subsequently, the PCR product was digested with XhoI and EcoRI and cloned into pMIG (pMIG was a
612 gift from William Hahn (Addgene plasmid #9044; http://n2t.net/addgene:9044; RRID:Addgene_9044). Retroviral
613 particles were produced by transient transfection in combination with packaging plasmid pUMVC (pUMVC was a gift
614 from Bob Weinberg (Addgene plasmid #8449)) and pseudotyping plasmid pMD2.G (pMD2.G was a gift from Didier
615 Trono (Addgene plasmid #12259; http://n2t.net/addgene:12259; RRID:Addgene_12259)) using TransIT-LT (Mirus).
616 Viral supernatants were collected 48-72 hours after transfection, clarified by centrifugation, supplemented with 10%
617 FCS and Polybrene (Sigma-Aldrich, final concentration of 8µg/ml) and 0.45µm filtered (Millipore; SLHP033RS). Cell
618 transduction was performed by incubating the PDGFß cells with viral supernatants for 48 hours. eGFP expressing
619 cells were single cell sorted.

620 ***RNA isolation, RNA-Seq library preparation and sequencing***

621 RNA was extracted according to the manufacturer s instructions using the RNeasy Kit (QIAGEN). For RNA-seq
622 Illumina TruSeq Stranded Total RNA Library Preparation Kit was used using 1000 ng RNA as input. Sequencing
623 libraries were quantified using Tapestation (Agilent) and Quantus (Promega). Equimolar pooling of the libraries was

normalized to 1.8 pM, denatured using 0.2 N NaOH and neutralized with 200 nM Tris pH 7.0 prior to sequencing. Final sequencing was performed on a NextSeq500/550 platform (Illumina) according to the manufacturer's protocols (Illumina, CA, USA).

Analysis of RNA-seq data

Pipeline nf-core/rnaseq[71] was used to analyse RNA-seq data. Briefly, reads were aligned to hg38 reference genome using STAR[72] and gene expression was quantified with Salmon[73]. Deferentially expressed genes were identified using DESeq2[74]. We used adjusted p-value of 1e-05 and log2 fold change of 1 as thresholds to select the significant DE genes, which were used as true labels to evaluate the Runx1 target gene prediction (see above). GO enrichment analysis was performed R package gprofiler2 and we showed results for biological process and pathways from Human Phenotype Ontology (**Supplementary Fig. 11e**). Volcano plot was generated by using R package EnhancedVolcano[75].

Calculation of population-doubling level (PDL)

For determining PDL, PDGFRb cells overexpressing Runx1 (or as control having genomicaly integrated the empty vector sequence) were passaged in 6-well plates at density of 1,5x 10(4) cells/well. Every 96hrs (at sub-confluent state), cells were harvested and counted in a hemocytometer before re-seeded at initial density.

Ethics

The ethics committee of the University Hospital RWTH Aachen approved the human tissue protocol for cell isolation (EK-016/17). Kidney tissues were collected from the Urology Department of the University Hospital Eschweiler from patients undergoing nephrectomy due to renal cell carcinoma.

Statistical analysis

All reported p-values based on multi-comparison tests were corrected using the Benjamini-Hochberg method.

Data availability

The scATAC-seq data generated from UUO mouse kidney and RNA-seq data from Runx1 overexpression of human fibroblasts have been deposited in NCBI's Gene Expression Omnibus and are accessible through GEO Series accession number GSE139950.

Code availability

The scOpen code is available at `https://github.com/CostaLab/scopen` and can be installed by *pip install scopen*. All scripts for reproducing the analysis are available at `https://github.com/CostaLab/scopen-reproducibili` as well as tables with all benchmarking results and raw count matrices from benchmarking datasets. Tutorial for the

use of HINT-ATAC with the hematopoetic data set is provided in https://www.regulatory-genomics.org/hint/tutorial-differential-footprints-on-scatac-seq/.

Acknowledgements

This work was funded by grants of the Interdisciplinary Center for Clinical Research (IZKF) Aachen, RWTH Aachen University Medical School, Aachen, Germany and by the Deutsche Forschungsgemeinschaft (DFG-GE 2811/3) to I.C. and (DFG SFB/TRR57 P30, SFB/TRR219 P5) and a Grant of the European Research Council (ERC-StG 677448) to R.K. and by the Bundesministerium für Bildung und Forschung (BMBF e:Med Consortia Fibromap) to I.C. and R.K.. C.K. was partly funded by the clinician scientist program of the German Society of Internal Medicine (DGIM) and a Gerok position of the DFG SFB/TRR 219, P5. Simulations were performed with computing resources granted by ITC RWTH Aachen University under project rwth0233 and rwth0429. We thank the team of the IZKF Aachen Genomics Core facility for sequencing experiments.

Author contributions

Z.L., I.C., C.K., R.K. conceived the experiments, Z.L., C.K., M.C., S.Z. and S.M. conducted the experiments. All authors analysed the results and reviewed the manuscript.

Competing interests

The authors declare no competing interests.

References

1. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. methods* **10**, 1213–1218 (2013).
2. Corces, M. R. *et al.* The chromatin accessibility landscape of primary human cancers. *Science* **362**, eaav1898 (2018).
3. Schep, A. N. *et al.* Structured nucleosome fingerprints enable high-resolution mapping of chromatin architecture within regulatory regions. *Genome Res.* **25**, 1757–1770 (2015).
4. Li, Z. *et al.* Identification of transcription factor binding sites using ATAC-seq. *Genome biology* **20**, 45 (2019).
5. Buenrostro, J. D. *et al.* Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523**, 486 (2015).

6. Buenrostro, J. D. *et al.* Integrated single-cell analysis maps the continuous regulatory landscape of human hematopoietic differentiation. *Cell* **173**, 1535–1548 (2018).

7. Domcke, S. *et al.* A human cell atlas of fetal chromatin accessibility. *Science* **370** (2020).

8. Satpathy, A. T. *et al.* Transcript-indexed ATAC-seq for precision immune profiling. *Nat. medicine* **24**, 580–590 (2018).

9. Zamanighomi, M. *et al.* Unsupervised clustering and epigenetic classification of single cells. *Nat. communications* **9**, 2410 (2018).

10. Bravo González-Blas, C. *et al.* cisTopic: cis-regulatory topic modeling on single-cell ATAC-seq data. *Nat. Methods* **16**, 397–400 (2019).

11. Fang, R. *et al.* Comprehensive analysis of single cell atac-seq data with snapatac. *Nat. communications* **12**, 1–15 (2021).

12. Schep, A. N., Wu, B., Buenrostro, J. D. & Greenleaf, W. J. chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. *Nat. methods* **14**, 975 (2017).

13. Pliner, H. A. *et al.* Cicero predicts cis-regulatory DNA interactions from single-cell chromatin accessibility data. *Mol. cell* **71**, 858–871 (2018).

14. Van Dijk, D. *et al.* Recovering gene interactions from single-cell data using data diffusion. *Cell* **174**, 716–729 (2018).

15. Gong, W., Kwak, I.-Y., Pota, P., Koyano-Nakagawa, N. & Garry, D. J. DrImpute: imputing dropout events in single cell RNA sequencing data. *BMC bioinformatics* **19**, 220 (2018).

16. Becht, E. *et al.* Dimensionality reduction for visualizing single-cell data using UMAP. *Nat. biotechnology* **37**, 38 (2019).

17. Xiong, L. *et al.* SCALE method for single-cell ATAC-seq analysis via latent feature extraction. *Nat. communications* **10**, 1–10 (2019).

18. Cichocki, A. & Phan, A.-H. Fast local algorithms for large scale nonnegative matrix and tensor factorizations. *IEICE transactions on fundamentals electronics, communications computer sciences* **92**, 708–721 (2009).

19. Satopaa, V., Albrecht, J., Irwin, D. & Raghavan, B. Finding a" kneedle" in a haystack: Detecting knee points in system behavior. In *2011 31st international conference on distributed computing systems workshops*, 166–171 (IEEE, 2011).

20. Chen, H. *et al.* Assessment of computational methods for the analysis of single-cell atac-seq data. *Genome biology* **20**, 1–25 (2019).

21. Hou, W., Ji, Z., Ji, H. & Hicks, S. C. A systematic evaluation of single-cell RNA-sequencing imputation methods. *Genome Biol.* **21**, 218 (2020). DOI 10.1186/s13059-020-02132-x.

22. Huang, M. *et al.* SAVER: gene expression recovery for single-cell RNA sequencing. *Nat. methods* **15**, 539 (2018).

23. Li, W. V. & Li, J. J. An accurate and robust imputation method scImpute for single-cell RNA-seq data. *Nat. communications* **9**, 997 (2018).

24. Eraslan, G., Simon, L. M., Mircea, M., Mueller, N. S. & Theis, F. J. Single-cell RNA-seq denoising using a deep count autoencoder. *Nat. communications* **10**, 390 (2019).

25. Li, R. & Quon, G. scBFA: modeling detection patterns to mitigate technical noise in large-scale single-cell genomics data. *Genome biology* **20**, 193 (2019).

26. Josse, J., Husson, F. *et al.* missmda: a package for handling missing values in multivariate data analysis. *J. Stat. Softw.* **70**, 1–31 (2016).

27. Hubert, L. & Arabie, P. Comparing partitions. *J. classification* **2**, 193–218 (1985).

28. Becht, E. *et al.* Dimensionality reduction for visualizing single-cell data using umap. *Nat. biotechnology* **37**, 38–44 (2019).

29. Kramann, R. *et al.* Pharmacological GLI2 inhibition prevents myofibroblast cell-cycle progression and reduces kidney fibrosis. *The J. clinical investigation* **125**, 2935–2951 (2015).

30. Kramann, R. *et al.* Perivascular Gli1+ progenitors are key contributors to injury-induced organ fibrosis. *Cell stem cell* **16**, 51–66 (2015).

31. Korsunsky, I. *et al.* Fast, sensitive and accurate integration of single-cell data with harmony. *Nat. methods* **16**, 1289–1296 (2019).

32. Wu, H., Kirita, Y., Donnelly, E. L. & Humphreys, B. D. Advantages of Single-Nucleus over Single-Cell RNA Sequencing of Adult Kidney: Rare Cell Types and Novel Cell States Revealed in Fibrosis. *J. Am. Soc. Nephrol.* **30**, 23–32 (2019).

33. Stuart, T. *et al.* Comprehensive integration of single-cell data. *Cell* **177**, 1888–1902 (2019).

34. Granja, J. M. *et al.* Archr is a scalable software package for integrative single-cell chromatin accessibility analysis. Tech. Rep., Nature Publishing Group (2021).

35. Bábíčková, J. *et al.* Regardless of etiology, progressive renal disease causes ultrastructural and functional alterations of peritubular capillaries. *Kidney international* **91**, 70–85 (2017).

36. Kramann, R. *et al.* Parabiosis and single-cell RNA sequencing reveal a limited contribution of monocytes to myofibroblasts in kidney fibrosis. *JCI insight* **3** (2018).

37. Vaidya, V. S., Ramirez, V., Ichimura, T., Bobadilla, N. A. & Bonventre, J. V. Urinary kidney injury molecule-1: a sensitive quantitative biomarker for early detection of kidney tubular injury. *Am. journal physiology. Ren. physiology* **290**, F517–29 (2006).

38. Sugawara, A., Sanno, N., Takahashi, N., Osamura, R. Y. & Abe, K. Retinoid X receptors in the kidney: their protein expression and functional significance. *Endocrinology* **138**, 3175–80 (1997).

39. Marable, S. S., Chung, E., Adam, M., Potter, S. S. & Park, J.-S. Hnf4a deletion in the mouse kidney phenocopies Fanconi renotubular syndrome. *JCI Insight* **3**, 354–80 (2018).

40. Kramann, R., DiRocco, D. P. & Humphreys, B. D. Understanding the origin, activation and regulation of matrix-producing myofibroblasts for treatment of fibrotic disease. *The J. pathology* **231**, 273–289 (2013).

41. Kuppe, C. *et al.* Decoding myofibroblast origins in human kidney fibrosis. *Nature* **589**, 281–286 (2021).

42. Muhl, L. *et al.* Single-cell analysis uncovers fibroblast heterogeneity and criteria for fibroblast and mural cell identification and discrimination. *Nat. Commun.* (2020).

43. de Bruijn, M. & Dzierzak, E. Runx transcription factors in the development and function of the definitive hematopoietic system. *Blood* **129**, 2061–2069 (2017).

44. Chan, S. C. *et al.* Mechanism of fibrosis in HNF1B-related autosomal dominant tubulointerstitial kidney disease. *J. Am. Soc. Nephrol.* (2018). DOI 10.1681/ASN.2018040437.

45. Henderson, N. C. *et al.* Targeting of αv integrin depletion identifies a core, targetable molecular pathway that regulates fibrosis across solid organs. *Nat. Medicine* (2013).

46. Buenrostro, J. D., Wu, B., Chang, H. Y. & Greenleaf, W. J. ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. *Curr. Protoc. Mol. Biol.* **109** (2015).

47. Andrews, T. S. & Hemberg, M. False signals induced by single-cell imputation. *F1000Research* **7**, 1740 (2018). DOI 10.12688/f1000research.16613.1.

48. Rotem, A. *et al.* Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state. *Nat. Biotechnol.* **33**, 1165–1172 (2015).

49. Bartosovic, M., Kabbe, M. & Castelo-Branco, G. Single-cell CUT&Tag profiles histone modifications and transcription factors in complex tissues. *Nat. Biotechnol.* (2021). DOI 10.1038/s41587-021-00869-9.

50. Kaya-Okur, H. S. *et al.* CUT&Tag for efficient epigenomic profiling of small samples and single cells. *Nat. Commun.* **10**, 1930 (2019).

51. Smallwood, S. A. *et al.* Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nat. Methods* **11**, 817–820 (2014).

52. Zhou, T. *et al.* Runt-Related Transcription Factor 1 (RUNX1) Promotes TGF-β-Induced Renal Tubular Epithelial-to-Mesenchymal Transition (EMT) and Renal Fibrosis through the PI3K Subunit p110δ. *EBioMedicine* (2018). DOI 10.1016/j.ebiom.2018.04.023.

53. Koth, J. *et al.* Runx1 promotes scar deposition and inhibits myocardial proliferation and survival during zebrafish heart regeneration. *Development* **147** (2020).

54. Salton, G. & McGill, M. J. Introduction to modern information retrieval. (1986).

55. Hsieh, C.-J. & Dhillon, I. S. Fast coordinate descent methods with variable selection for non-negative matrix factorization. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, 1064–1072 (2011).

56. Corces, M. R. *et al.* Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. *Nat. genetics* **48**, 1193–1203 (2016).

57. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal* **17**, 10–12 (2011).

58. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. methods* **9**, 357 (2012).

59. Institute, B. Picard tools. http://broadinstitute.github.io/picard/ (2019). Accessed: 2019-01-01; version 2.18.22.

60. Li, H. *et al.* The sequence alignment/map format and samtools. *Bioinformatics* **25**, 2078–2079 (2009).

61. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome biology* **9**, R137 (2008).

62. Davis, J. & Goadrich, M. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning*, 233–240 (2006).

63. Raimundo, F., Vallot, C. & Vert, J.-P. Tuning parameters of dimensionality reduction methods for single-cell rna-seq analysis. *Genome biology* **21**, 1–17 (2020).

64. Maaten, L. v. d. & Hinton, G. Visualizing data using t-SNE. *J. machine learning research* **9**, 2579–2605 (2008).

65. Cusanovich, D. A. *et al.* A single-cell atlas of in vivo mammalian chromatin accessibility. *Cell* **174**, 1309–1324 (2018).

66. Khan, A. *et al.* JASPAR 2018: Update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic Acids Res.* **46**, D260–D266 (2018).

67. Cairns, J. *et al.* Chicago: robust detection of dna looping interactions in capture hi-c data. *Genome Biol.* **17**, 127 (2016).

68. Grau, J., Grosse, I. & Keilwagen, J. Prroc: computing and visualizing precision-recall and receiver operating characteristic curves in r. *Bioinformatics* **31**, 2595–2597 (2015).

69. Fornes, O. *et al.* Jaspar 2020: update of the open-access database of transcription factor binding profiles. *Nucleic acids research* 1 (2019).

70. Angerer, P. *et al.* destiny: diffusion maps for large-scale single-cell data in r. *Bioinformatics* **32**, 1241–1243 (2016).

71. Patel, H. *et al.* nf-core/rnaseq: nf-core/rnaseq v3.0 - silver shark (2020). URL https://doi.org/10.5281/zenodo.4323183. DOI 10.5281/zenodo.4323183.

72. Dobin, A. *et al.* Star: ultrafast universal rna-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

73. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. methods* **14**, 417–419 (2017).

74. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology* **15**, 1–21 (2014).

75. Blighe, K., Rana, S. & Lewis, M. Enhancedvolcano: Publication-ready volcano plots with enhanced colouring and labeling. *R package version* **1** (2019).